

EXHIBIT 1

**UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK**

AUTHORS GUILD, et al.,

Plaintiffs,

v.

OPEN AI INC., et al.,

Defendants.

JONATHAN ALTER, et al.,

Plaintiffs,

v.

OPENAI, INC., et al.,

Defendants.

ECF CASE

No. 1:23-cv-08292-SHS;

No. 1:23-cv-10211-SHS

**PLAINTIFFS' SECOND SET OF
REQUESTS FOR PRODUCTION
TO MICROSOFT
CORPORATION**

PLAINTIFFS' SECOND SET OF REQUESTS FOR PRODUCTION TO DEFENDANT

MICROSOFT CORPORATION

Pursuant to Rules 26 and 34 of the Federal Rules of Civil Procedure, Plaintiffs request that Defendant Microsoft Corporation ("Microsoft") respond to the following Second Set of Requests for Production of Documents ("Requests"). Responses to these Requests shall be due within thirty (30) days of the date of service or as otherwise mutually agreed by the parties. Plaintiffs are amenable to an electronic production, subject to agreement by the parties.

In accordance with Rule 34(b), Microsoft shall provide written responses to the following Requests and shall produce the requested documents as they are kept in the ordinary and usual course of business or shall organize and label the documents to correspond with the categories in

these Requests. In accordance with Rule 26(e), Microsoft shall supplement or correct its responses or productions as necessary.

I. DEFINITIONS

1. **“Actions”** means the above captioned litigation, *Authors Guild et al. v. Open AI Inc. et al.*, No. 1:23-cv-08292-SHS (S.D.N.Y.) (opened Sept. 19, 2023 and amended December 5, 2023) (“Fiction Docket”), and *Alter et al. v. OpenAI et al.*, No / 1:23-cv-10211-SHS (S.D.N.Y.) (opened Nov. 21, 2023, and amended Dec. 19, 2023) (“Nonfiction Docket”).

2. **“Class Works”** pertains to both Fiction Class Works and Nonfiction Class Works as defined in ¶394 and in ¶397 of the First Consolidated Complaint (Fiction Docket No. 69).

3. **“Communication(s)”** means the transmittal of information (in the form of facts, ideas, inquiries or otherwise) by any means, including, but not limited to, telephone calls, emails (whether via company server or personal webmail or similar accounts), faxes, text messages (on work or personal phones), instant messages, Skype, Line, WhatsApp, WeChat, other electronic messages, letters, notes, and voicemails.

4. **“Document(s)”** is defined to be synonymous in meaning and equal in scope to the usage of the term “documents or electronically stored information” in Rule 34(a)(1)(A). For the avoidance of doubt, **Document(s)** includes **Communication(s)**.

5. **“You”, “Your”, and “Microsoft”** means Microsoft Corporation and any of its directors, officers, employees, partners, members, representatives, agents (INCLUDING attorneys, accountants, consultants, investment advisors or bankers), and any other person acting or purporting to act on their behalf, as well as corporate parents, subsidiaries, affiliates, predecessor entities, successor entities, divisions, departments, groups, acquired entities, related entities, or any other entity acting or purporting to act on its behalf.

6. **“OpenAI”** means OpenAI, Inc., OpenAI GP, LLC, OpenAI, LLC, OpenAI OPCO LLC, OpenAI Global LLC, OAI Corporation, LLC, OpenAI Holdings, and any of their directors, officers, employees, partners, members, representatives, agents (including attorneys, accountants, consultants, investment advisors or bankers), and any other person acting or purporting to act on their behalf, as well as corporate parents, subsidiaries, affiliates, predecessor entities, successor entities, divisions, departments, groups, acquired entities, related entities, or any other entity acting or purporting to act on their behalf.

7. **“Large Language Model,” “LLM,” “AI Model(s),” “Generative AI system(s),” “model(s),” and “API Product(s)”** have the same meaning as they are used in OpenAI’s letter to the Register of Copyrights and Director of the U.S. Copyright Office dated October 30, 2023, “Re: Notice of Inquiry and Request for Comment [Docket No. 2023-06]” and include all models listed or described in <https://platform.openai.com/docs/models>.

8. **“Person(s)”** means any individual or entity.

9. **“Fine Tune(d),” “Fine Tuning,” “Pre-Train(ed),” “Pre-Training,” “Train[ed],” and “Training,”** have the same meaning as the term is discussed in **Your** website materials and statements. See e.g., <https://platform.openai.com/docs/guides/fine-tuning> (“OpenAI’s text generation models have been pre-trained on a vast amount of text. To use the models effectively, we include instructions and sometimes several examples in a prompt. Using demonstrations to show how to perform a task is often called “few-shot learning.” Fine-tuning improves on few-shot learning by training on many more examples than can fit in the prompt, letting you achieve better results on a wide number of tasks. Once a model has been fine-tuned, you won’t need to provide as many examples in the prompt.”); <https://openai.com/index/language-unsupervised/> (“These results provide a convincing example that pairing supervised learning methods with unsupervised

pre-training works very well; this is an idea that many have explored in the past, and we hope our result motivates further research into applying this idea on larger and more diverse datasets.”); and <https://openai.com/index/how-should-ai-systems-behave/> (“The two main steps involved in building ChatGPT work as follows . . . First, we “pre-train” models by having them predict what comes next in a big dataset . . . Then, we “fine-tune” these models on a more narrow dataset that we carefully generate with human reviewers who follow guidelines that we provide them.”); <https://openai.com/index/gpt-4-research/> (“Interestingly, the base pre-trained model is highly calibrated . . . Note that the model’s capabilities seem to come primarily from the pre-training process . . . ”); <https://openai.com/index/text-and-code-embeddings-by-contrastive-pre-training/> (“In this work, we show that contrastive pre-training on unsupervised data at scale leads to high quality vector representations of text and code.”); <https://openai.com/index/vpt/> (“We trained a neural network to play Minecraft by Video PreTraining (VPT) on a massive unlabeled video dataset of human Minecraft play, while using only a small amount of labeled contractor data. With fine-tuning, our model can learn to craft diamond tools . . .”).

10. **“ChatGPT”** means all consumer-facing versions of the chatbot application OpenAI released in November 2022 (and discussed in paragraphs 83-95 of the First Consolidated Complaint (Fiction Docket No. 69)) and any LLM underlying those consumer-facing applications.

11. **“Concern”** means be the subject of, make reference to, comment on, discuss, describe, identify, or contain text or images about the stated topic.

12. **“Including”** means including but not limited to.

13. **“Investors”** means any person or entity that financially invested in the operations and/or entitie(s) controlled and/or overseen by one or more of the OpenAI defendant entities.

14. **“Relate to”** means refer to, respond to, describe, evidence, or constitute, in whole or in part.

15. **“Publishers”** means a person or entity whose business is publishing materials, including, but not limited to, books, periodicals, magazines, and newspapers.

16. **“Training Material”** means any data (e.g. text) obtained for or used to train a Large Language Model.

II. RELEVANT TIME PERIOD

The relevant time period is January 1, 2015 through the present (“Relevant Time Period”), unless otherwise specifically indicated, and shall include all **Documents** and any other information relating to such period, even though prepared or published outside of the Relevant Time Period. If a **Document** prepared before the Relevant Time Period is necessary for a correct or complete understanding of any **Document** covered by any of these Requests, please provide the earlier **Document** as well. If any **Document** is undated and the date of its preparation cannot be determined, please produce the **Document** if it is otherwise responsive to any Request.

III. INSTRUCTIONS

1. The production by one person, party, or entity of a **Document** does not relieve another person, party, or entity from the obligation to produce his, her, or its own copy of that **Document**.

2. Produce **Documents** not otherwise responsive to these Requests if such **Documents** relate to the **Documents** that are called for by these Requests, or if such **Documents** are attached to **Documents** called for by these Requests.

3. Produce each **Document** requested herein in its entirety and without deletion, excisions, redaction, or other modification regardless of whether YOU consider the entire document to be relevant or responsive.

4. If any **Document** is known to have existed but no longer exists, has been destroyed, or is otherwise unavailable, identify the **Document**, the reason for its loss, destruction, or unavailability, the name of each person known or reasonably believed by Microsoft to have had possession, custody, or control of the original and any copy thereof (if applicable), and a description of the disposition of each copy of the **Document**.

5. If no **Documents** responsive to a particular Request exist, state that no responsive **Documents** exist.

6. If You assert that any of the **Documents** and things requested are protected from discovery by attorney-client privilege, the attorney work product doctrine, or any other evidentiary privilege, specify for each **Document** (1) the grounds asserted as the reason for non-production; (2) the date the **Document** was prepared; (3) the identity of the attorney(s) who drafted or received the **Document(s)** (if attorney-client privilege or attorney work product protection is claimed); (4) the identity of the parties who prepared or received the **Document**; and (5) the nature of the **Document**.

7. Construe the conjunctions “and” and “or” non-restrictively or non-exclusively if doing so would bring within the scope of these Requests **Documents** that might otherwise be construed to be outside of their scope.

8. Construe the use of the singular to include the plural; the use of the masculine, feminine, or neuter gender to include the others; and the use of one form of the verb to include the others if doing so would bring within the scope of these Requests **Documents** that might otherwise be construed to be outside of their scope.

Plaintiffs, by and through their undersigned attorneys, request that Defendants provide a response to these Requests within thirty (30) days of the date of service hereof as provided by Federal Rule of Civil Procedure 33.

REQUESTS FOR PRODUCTION

REQUEST FOR PRODUCTION NO. 41:

Documents sufficient to identify all datasets containing commercial works of fiction and/or nonfiction, including Class Works, that **You** or **OpenAI** accessed, downloaded or copied to train **Large Language Models**.

REQUEST FOR PRODUCTION NO. 42:

Documents Concerning or Relating to any complaints **You** have received regarding **Your** or **OpenAI's** use of potentially copyrighted material in training **Large Language Models**.

REQUEST FOR PRODUCTION NO. 43:

Documents Concerning or Relating to Your policies or procedures **Relating to** receiving, processing, or responding to complaints regarding **Your** or **OpenAI's** use of potentially copyrighted material in training **Large Language Models**.

REQUEST FOR PRODUCTION NO. 44:

Documents sufficient to identify any effort by **You** or **OpenAI** to remedy or mitigate **Your** or **OpenAI's** unauthorized use of copyrighted material to train any **Large Language Models**.

REQUEST FOR PRODUCTION NO. 45:

All **Communications** between January 1, 2016 and December 31, 2022, related to the use or reproduction of copyrighted materials to train any artificial intelligence model **OpenAI** developed, including without limitation all **Large Language Models** and diffusion models (e.g., DALL-E, Codex, GPT-2).

REQUEST FOR PRODUCTION NO. 46:

Documents sufficient to show all attempts by **You** or **OpenAI** to obtain a license for the use of text or audio works in connection with the development of any **Large Language Model(s)**, including but not limited to **Documents** reflecting any negotiations of such licenses.

REQUEST FOR PRODUCTION NO. 47:

All **Documents Concerning** or **Relating to Your** decision(s) to seek or obtain a license to use any copyrighted material as **Training Material**.

REQUEST FOR PRODUCTION NO. 48:

All **Documents Concerning** or **Relating to** the relative value of different types of **Training Material**.

REQUEST FOR PRODUCTION NO. 49:

All **Documents** concerning or relating to **OpenAI** deriving **Training Material** from any physical books or audiobooks.

REQUEST FOR PRODUCTION NO. 50:

All **Documents Concerning** or **Relating to** OpenAI's use of digital books as **Training Material**.

REQUEST FOR PRODUCTION NO. 51:

Documents sufficient to show any reproduction of the data **OpenAI** used to train any **Large Language Model**, including any reproductions that came into Microsoft's possession.

REQUEST FOR PRODUCTION NO. 52:

All **Documents** and **Communications Concerning** or **Relating** to the use or contemplated use of material from any text data repository (e.g. Library Genesis, Internet Archive, Z-Library,

Common Crawl, WebText, Project Gutenberg, Anna’s Archive, Open Library, Reddit, DuXiu) as **Training Material**.

REQUEST FOR PRODUCTION NO. 53:

All **Documents Concerning or Relating to OpenAI’s** potential exhaustion of available text data for use as **Training Material**.

REQUEST FOR PRODUCTION NO. 54:

Documents sufficient to show when **You** became aware of each dataset used to train **OpenAI’s Large Language Models**, including but not limited to, all **Documents Concerning and Relating to** the “writeup of OpenAI’s technical results on large language models” **OpenAI** sent to **You** in February 2020 that “include[d] details on training methods, datasets, etc to facilitate comparisons and collaboration.” (MSFT_AICPY_000228653).

REQUEST FOR PRODUCTION NO. 55:

All **Documents Concerning or Relating to** the relative value of different types of text data in training **Large Language Models**.

REQUEST FOR PRODUCTION NO. 56:

All **Documents Concerning or Relating to** any programming of any **Large Language Models** developed by **OpenAI** designed or intended—in whole or in part—to avoid copyright infringement and/or to avoid reproducing materials used to train the **Large Language Models** or **ChatGPT**.

REQUEST FOR PRODUCTION NO. 57:

All **Documents Concerning or Relating to** the impact that including or excluding human-generated text from a **Large Language Model’s** training data has on the quality of its outputs.

REQUEST FOR PRODUCTION NO. 58:

Documents sufficient to show any exclusive or non-exclusive rights granted to **You** by **OpenAI** to commercialize, monetize, or productize **OpenAI's Large Language Models**, training data, or intellectual property.

REQUEST FOR PRODUCTION NO. 59:

Documents sufficient to identify all of Your commercial products that use or include **any Large Language Model developed by OpenAI** , including but not limited to, **Your** Bing search engine (as set forth in the First Consolidated Complaint [Fiction Docket No. 69] at ¶166) as well as Microsoft Teams, Microsoft 365 Copilot and GitHub Copilot (*id.* at ¶167).

REQUEST FOR PRODUCTION NO. 60:

Documents sufficient to identify each person within **Microsoft** who participated, was involved in or was consulted **Concerning** the use, implementation or incorporation of **ChatGPT** or **Large Language Models** in **Your** commercial products.

REQUEST FOR PRODUCTION NO. 61:

All **Documents Concerning** or **Relating to** customers substituting away from or towards a **Microsoft** product using a **Large Language Model** based on the quality of the **Large Language Model**.

REQUEST FOR PRODUCTION NO. 62:

Documents sufficient to identify **Your** gross revenues, net revenues, and profits, by month, from Your commercial products that include or use **ChatGPT** or **Large Language Models**.

REQUEST FOR PRODUCTION NO. 63:

Documents sufficient to show the gross revenues, net revenues, and profits, by month, generated by each **Large Language Model You** have commercialized, sold, and/or licensed, either as a standalone product or as part of another product.

REQUEST FOR PRODUCTION NO. 64:

Document sufficient to show Your balance sheet, income statement, and cash flow statement, on a monthly basis during the relevant time period.

REQUEST FOR PRODUCTION NO. 65:

All **Documents Concerning or Relating to** predictions, forecasts, or projections of profits, revenues or cash flows of **OpenAI** or from **Large Language Models** or products using or containing **Large Language Models**.

REQUEST FOR PRODUCTION NO. 66:

All presentations, memoranda, or other **Documents** provided to **Your** Board of Directors **Concerning or Relating to OpenAI's** development of Large Language Models.

REQUEST FOR PRODUCTION NO. 67:

Documents sufficient to show each version of the terms and/or conditions of use of **Your** products that use or include **ChatGPT** or **Large Language Models**.

REQUEST FOR PRODUCTION NO. 68:

Retention agreements You entered from January 1, 2016 to the present, with any law firm(s) for actual or potential copyright litigation **Related to Your** investment in and/or partnership with **OpenAI**.

REQUEST FOR PRODUCTION NO. 69:

Documents sufficient to show all indemnification provisions agreed to by and between **You** and **OpenAI**.

REQUEST FOR PRODUCTION NO. 70:

All **Documents Concerning** or **Relating to** whether **OpenAI's** use of copyrighted material as Training Material complies with copyright law in the United States or elsewhere.

REQUEST FOR PRODUCTION NO. 71:

All **Documents You** received from OpenAI in the course of **Your** due diligence regarding each of **Your** investments in **OpenAI**.

REQUEST FOR PRODUCTION NO. 72:

Documents and **Communications Concerning** or **Relating to Your** decision to modify **Your** normal due diligence process for the purposes of **Your** investment in **OpenAI**.

REQUEST FOR PRODUCTION NO. 73:

Documents sufficient to identify all data that was shared among **You** and **OpenAI** in the course of **Your** partnership with **OpenAI**, including but not limited to data shared pursuant to the terms of the Joint Development and Collaboration Agreement (the "JDCA") (initial JDCA produced at MSFT_AICPY_000000345) since June 2019.

REQUEST FOR PRODUCTION NO. 74:

Documents sufficient to identify each of **Your** employees who at any point had access to "Restricted Access Model Training Code" pursuant to Section 3(e) of the Second Amended and Restated JDCA (the "Second Amended JDCA") (MSFT_AICPY_000000575).

REQUEST FOR PRODUCTION NO. 75:

Documents sufficient to identify any “secure repository” **You** identified to **OpenAI** pursuant to Section 3(e)(iv) the Amended and Restated JDCA (MSFT_AICPY_000004210).

REQUEST FOR PRODUCTION NO. 76:

Documents sufficient to identify all individuals involved in copying the “Development Materials”—as defined in the Second Amended JDCA—for storage in **Your** secure repository, as set forth in Section 3(h)(iv) the Second Amended JDCA.

REQUEST FOR PRODUCTION NO. 77:

All **Documents Concerning or Relating to** the Data Working Group, as referenced in the Second Amended JDCA, including but not limited to all **Communications** among members of, meeting minutes of, and data policies compiled by the Data Working Group.

REQUEST FOR PRODUCTION NO. 78:

All **Documents** created in the course of the Governing Board’s—as referenced in the Second Amended JDCA—duties, including but not limited to meeting minutes, memorandum and **Communications**.

Dated: August 14, 2024

Respectfully submitted,

/s/ Rachel Geman

Rachel Geman
Wesley Dozier (*pro hac vice*)
Anna J. Freymann
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
250 Hudson Street, 8th Floor
New York, NY 10013-1413
Telephone: 212.355.9500
rgeman@lchb.com
wdozier@lchb.com
afreymann@lchb.com

Reilly T. Stoler (*pro hac vice*)
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
275 Battery Street, 29th Floor
San Francisco, CA 94111-3339
Telephone: 415.956.1000
rstoler@lchb.com

/s/ Justin Nelson

Justin A. Nelson (*pro hac vice*)
Alejandra C. Salinas (*pro hac vice*)
SUSMAN GODFREY L.L.P.
1000 Louisiana Street, Suite 5100
Houston, TX 77002
Telephone: 713-651-9366
jnelson@susmangodfrey.com
asalinas@susmangodfrey.com

Rohit D. Nath (*pro hac vice*)
SUSMAN GODFREY L.L.P.
1900 Avenue of the Stars, Suite 1400
Los Angeles, California 90067
Telephone: 310-789-3100
rnath@susmangodfrey.com

J. Craig Smyser
SUSMAN GODFREY L.L.P.
1901 Avenue of the Americas, 32nd Floor
New York, New York 10019
Telephone: 212-336-8330
csmyser@susmangodfrey.com

/s/ Scott Sholder

Scott J. Sholder

CeCe M. Cole

COWAN DEBAETS ABRAHAMS & SHEPPARD LLP

41 Madison Avenue, 38th Floor

New York, New York 10010

Telephone: 212.974.7474

ssholder@cdas.com

ccole@cdas.com

Attorneys for Plaintiffs and the Proposed Classes

CERTIFICATE OF SERVICE

I hereby certify that on August 14, 2024, a copy of the foregoing was served via electronic mail to all counsel of record in this matter.

/s/ Ariana Delucchi
(Signature)